Faculty of Engineering and Information Technology
University of Technology Sydney

High-density Visual Crowd Counting with Perspective Understanding in Deep Neural Networks

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Muming Zhao

January 2020

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by the Australian Government Research Training Program.

This thesis is the result of me conducted jointly with Shanghai Jiao Tong University as part of a collaborative Doctoral degree.

Signature of Candidate: Production Note: Signature removed prior to publication.

Date: 2020/01/23

Acknowledgments

I would like to thank my principal advisor Jian Zhang in UTS, for his continuous support either in my research or in the finance. He had always expressed his patience to me and never gave me up even at the time when I stuck in my progress. His enthusiasm and keen sense to the research helped guide and push me to complete different works in this thesis. I also want to express my gratitude to my co-supervisor Chongyang Zhang in SJTU for his significant support of my research. He used to spend a lot of time to help me revise my paper and give me valuable comments to improve my academic skills. I also would like to thank my principal supervisor Wenjun Zhang in SJTU, for his continuous support of my dual PhD study in UTS and SJTU. Without the three supervisors, this thesis would be impossible.

I want to thank all my colleagues: Xiaoshui Huang, Yazhou Yao, Junjie Zhang, Jiangchao Yao and Yuangang Pan. I appreciate the time they have spent discussing with me, where I have got inspired a lot.

Finally and most essentially, I am grateful for all the support from my parents, my sister and my dear friends. They are the source of my strength.

Contents

List of Figures

List of Tables

List of Publications

Papers published

- **Muming Zhao**, Jian Zhang, Chongyang Zhang, Wenjun Zhang: Leveraging Heterogeneous Auxiliary Tasks to Assist Crowd Counting, *30th IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019.

- **Muming Zhao**, Jian Zhang, Chongyang Zhang, Wenjun Zhang: Towards Locally Consistent Object Counting with Constrained Multi-stage Convolutional Neural Networks, *14th Asian Conference on Computer Vision* (ACCV), 2018.

- **Muming Zhao**, Jian Zhang, Fatih Porikli, Chongyang Zhang, Wenjun Zhang: Learning a perspective-embedded deconvolution network for crowd counting, *IEEE International Conference on Multimedia and Expo* (ICME) , 2017.

Papers in submission

- **Muming Zhao**, Jian Zhang, Chongyang Zhang, Fatih Porikli, Bingbing Ni, Wenjun Zhang: Scale-aware Crowd Counting via Depth-embedded Convolutional Neural Networks, *Transactions on Circuits and Systems for Video Technology* (TCSVT), *under review*.

Abstract

With population growth and worldwide urbanization, crowd gathering in public places has become more common. Thus estimating the number of people and measuring their density has become essential for practical applications such as physical security control and public space management. However, the complex environments of crowded scenes have imposed several challenges to general counting algorithms, among which scale variations of pedestrians is one of the most significant problems. With varying-sized objects, it is rather difficult for density-based counting systems to generate appropriate density estimations that conform to scale variations, which usually significantly degrades the counting accuracy. To handle the perspective distortion and the related scale-variation problem, traditional methods mainly perform feature normalization for perspective correction. However, within the deep learning framework, the perspective distortion has not been explicitly considered and addressed. Can we extend the mechanism of perspective handling with the powerful deep learning technique for further improvement? In this dissertation, we focus on measuring crowd density through deep architectures with in-network perspective understanding. Three works are presented. First, we develop a depth-embedded network that augments the original features to be scale-aware for more accurate density estimation. The depth map of a scene is encoded, rectified and finally embedded into the network via a proposed depth embedding module. Thus the objects, although in the same class, will attain distinct representations according to their scales in the feature space, which will directly benefit scale-aware density estima-

tions. We include a comprehensive comparison with various state-of-the-art methods for the task of crowd counting to verify the efficacy of incorporating geometric priors. Second, a multi-stage model with region-based supervisions is constructed to obtain robust features with implicit understandings of the scene geometry. With the internal multi-stage learning mechanism, features could be refined and adjusted repeatedly to perceive the scale variations. Besides, with local-based supervisions, the model is further constrained to generate locally consistent densities that conform to object scale variations. Experiments are presented to validate the effectiveness of the proposed model for crowd counting. Third, we build a multi-task framework that drives the network to embed desired semantic/geometric/numeric attributes to handle various type of challenges for crowd counting. With the multi-fold regularization effects introduced by three auxiliary tasks, the intermediate features are driven to convey desired properties and thus help improve the main task of density estimation. Extensive experiments have been conducted to indicate the effectiveness of the proposed method.